Chunking with Max-Margin Markov Networks*

Tang Buzhou, Wang Xuan , and Wang Xiaolong

Department of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, 518055,China

{tangbuzhou5125@gmail.com, wangxuan@insun.hit.edu.cn, xiaolongwang@insun.hit.edu.cn}

Abstract. In this paper, we apply Max-Margin Markov Networks (M3Ns) to English base phrases chunking, which is a large margin approach combining both the advantages of graphical models(such as Conditional Random Fields, CRFs) and kernel-based approaches (such as Support Vector Machines, SVMs) to solve the problems of multi-label multi-class supervised classification. To show the efficiency of M3Ns, we compare it with CRFs and other relative systems on the data set of CoNLL-2000 comprehensively. The experiment results show that M3Ns achieves state-of-the-art performance with strong generalization ability, which is better than CRFs.

Keywords: max-margin markov networks; graphical models; conditional random fields; support vector machines; generalization ability

1. Introduction

Text chunking is an intermediate step towards full parsing, which consists of dividing a text in syntactically correlated parts of words. Tasks of chunking are extracting the non-overlapping segments from a stream of data and identifying them with non-recursive cores of various types of phrases. It can be solved as sequential labeling.

Many probabilistic graphical models such as Hidden Markov Models (HMMs) (Zhou et al.,2000; Sang and Buchholz, 2000), Maximum Entropy Models (MEs) (Koeling, 2000), Conditional Random Fields (CRFs) (Lafferty et al., 2001), Semi-Markov Random Fields (Collins, 2002a) have been applied to chunking for their abilities to deal with structured data by taking advantages of the potential of interactions in a factored way (Jordan al., 1998). However, the condition of the probabilistic infinite samples assumption cannot be satisfied in practice and over fitting problem cannot be avoided.

On the other hand, the tasks of chunking can be recognized as a classifying problem, statistic machine learning techniques are also often applied to chunking and various machine learning approaches have been proposed for chunking such as SVMs (Cortes and Vapnik, 1995; Vapnik, 1999) and Boosting (Freund and Schapire, 1996). Compared with probabilistic graphical models, statistic machine learning approaches have no flaw of infinite samples assumption and have strong generalization guarantees theoretically, but they assume that the classification of each object (word or phrase) is independent and ignore some precious correlation information in structured data.

* This research has been partially supported by the National Natural Science Foundation of China (No.60435020 and No.90612005) and the Goal-oriented Lessons from the National 863 Program of China (No.2006AA01Z197).

M3Ns is a new framework combining SVMs with graphical models. It is a SVM-like approach that could also deal with structured data efficiently like graphical models do (Taskar, 2003). In practice, M3Ns can not only make use of correlations in structured data, including sequential data, like CRFs, but also efficiently deal with high-dimensional features with high generalization performance like SVMs.

In this paper, we apply M3Ns to the CoNLL-2000 text chunking shared task using distinct chunk representations. In addition, in order to investigate the generalization ability, we compare the performance of M3Ns and CRFs on data sets of different sizes.

2. Max-Margin Markov Networks

In statistical machine learning theory, the task is to learn a function h: X→Y from a training set of m i.i.d. samples $S = \{(x^i, y^i = t^i = t(x^i)) \in X \times Y \mid i = 1,..,m\}$, drawn from a fixed distribution Dx×y. The determinative function h is usual a linear function of features fj with coefficients w_j such that:

$$h_w(x) = \arg\max_y \sum_{j=1}^{n} w_j f_j(x, y) = \arg\max_y w^T f(x, y) \tag{1}$$

where n is feature space size.

For the sequence labeling problem, the data comes from a domain X×Y where X is a set and $Y=Y_1\times Y_2\times\ldots Y_k$ is a Cartesian product of the set of $Y_j=\{1,2,\ldots n_c\}$,j=1,…,k. Be different from most common classification setting, Y is not a single label, but a joint label for an whole sequence.

According to basis SVMs framework, the formal representation of the sequence label problem is provided as follows:

$$\min_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$
$$s.t. \quad w^T \Delta f_i(y) \geq l(y_i) - \varepsilon_i, \forall i, y. \tag{2}$$

Where $\Delta f_i(y) = f(x^i, y^i) - f(x^i, y)$ and $l_i(y)$ (called loss function), and the equivalent dual problem is:

$$\max_\alpha \sum_{i,y} \alpha_i(y) l_i(y) - \frac{1}{2} C \left\| \sum_{i,y} \alpha_i(y) \Delta f_i(y) \right\|^2$$
$$s.t. \quad \sum_{i,y} \alpha_i(y) = C, \forall i; \quad \alpha_i(y) \geq 0, \forall i \ , y. \tag{3}$$

In sequence labeling problems, the loss function can be various, such as per-label loss and the proportion of incorrect labels predicted. Here, per-label loss function is used in experiments.

Both the number of constraints in the primal QP in (2) and the number of variables in the dual QP (3) are exponential in the number of labels n_c. They can not be solved by general approaches. In M3Ns, the marginal dual variables are introduced as follows:

$$\mu_i(y_t, y_{t+1}) = \sum_{y \sim y_t y_{t+1}} \alpha_i(y), \forall i, y; \ \mu_i(y_t) = \sum_{y \sim y_t} \alpha_i(y), \forall i, y. \tag{4}$$

Where $y \sim y_t y_{t+1}$ denotes the full assignment y consistent with partial assignment: $y_t y_{t+1}$. In addition, the marginal dual variables must keep consistent between the pairs and singleton marginal:

$$\mu_i(y_{t+1}) = \sum_{y_t} \mu_i(y_t, y_{t+1}), \forall i, y. \tag{5}$$

Now, we can reformulate QP (3) in terms of these dual variables, and the original dual can be factored as follows:

$$\max_{\alpha} \sum_{i,y_t,y_{t+1}} \mu_i(y_t,y_{t+1}) l_i(y_t,y_{t+1}) - \frac{1}{2} C \left\| \sum_{i,y_t} \mu_i(y_t) \Delta f_i(y_t) \right\|^2$$
$$\sum_{i,y_t y_{t+1}} \mu_i(y_t,y_{t+1}) = C; \ \mu_i(y_t,y_{t+1}) > 0 \ \forall i,y. \quad \mu_i(y_{t+1}) = \sum_{y_t} \mu_i(y_t,y_{t+1}), \forall i,y. \tag{6}$$

Where the number of variables $\mu_i(y_t,y_{t+1})$ is O(mn_c^2).

To solve the problem (6), Taskar supplied a new SMO and ES algorithms and showed a generalization bound for the task of sequential labeling (Taskar, 2003). In our experiments, we adopt SMO and linear kernel.

3. Chunking

3.1.Chunking Representation

There is commonly one type of representation for text chunking — Inside/Outside representation. In order to describe the chunking more precisely, Uchimotoetal proposes a new representation for Japanese named entity extraction task (Uchimotoetal., 2000), and Xue introduces another new representation for Chinese segmentation task (Xue, 2003). We called them as Start/End representation. These two types are mentioned in (Taku Kudo and Yuji Matsumoto, 2000). In this paper, a new Start/End presentation will be introduced into chunking.

1、Inside/Outside

This representation uses the following set of three tags for representing proper chunks (Ramshaw and Marcus, 1995).

I Current token is inside of a chunk.
O Current token is outside of any chunk.
B Current token is the beginning of a chunk which immediately follows another chunk.

Tjong Kim Sang calls this method as IOB1 representation, and introduces three alternative versions — IOB2, IOE1and IOE2 (Tjong Kim Sang and Veenstra, 1999).

IOB2 A B tag is given for every token at the beginning of a chunk. Other tokens are the same as IOB1.

IOE1 An E tag is used to mark the last token of a chunk immediately preceding another chunk.

IOE2 An E tag is given for every token at the end of a chunk.

2、Start/End

This representation was first introduced in (Uchimotoetal., 2000), and is used for the Japanese named entity extraction task. It requires the following five tags for representing proper chunks.

B Current token is the start of a chunk consisting of more than one token.
E Current token is the end of a chunk consisting of more than one token.
I Current token is a middle of a chunk consisting of more than two tokens.
S Current token is a chunk consisting of only one token.
O Current token is outside of any chunk.

We called this representation as IOBES1 for convenience. Another representation was introduced in (Xue, 2003), and is used for the Chinese segmentation task. This method, called IOBES2, introduces two additional tags (B2 and B3) based on IOBES1 for chunks consisting of more than three tokens.

B2 A B2 tag is used to mark the first token immediately following B of a chunk consisting of more than three tokens.

B3 A B3 tag is used to mark the first token immediately following B2 of a chunk consisting of more than four tokens.

Similarly, we introduce another two tags (E2, E3) for chunks consisting more than three tokens.

E2 A E2 tag is used to mark the first token immediately preceding E of a chunk consisting of more than three tokens.

E3 A E3 tag is used to mark the first token immediately preceding E2 of a chunk consisting of more than four tokens.

We called this representation as IOBES3. In the CONLL-2000 text chunking shared task, the grammatical class of each chunk should be identified as a grammatical class label, and we represent them by a pair of an {I, O, B, E, S} label and a grammatical label. Examples of these representations of each phrase are shown in Table 1.

Table 1: Example for each chunk representation

	Inside/Outside				**Start/End**		
	IOB1	IOB2	IOE1	IOE2	IOBES1	IOBES2	IOBES3
He PRP	I-NP	B-NP	E-NP	E-NP	S-NP	S-NP	S-NP
reckons BZ	B-VP	B-VP	E-VP	E-VP	S-VP	S-VP	S-VP
the DT	B-NP	B-NP	I-NP	I-NP	B-VP	B-VP	B-VP
current JJ	I-NP	I-NP	I-NP	I-NP	I-VP	B2-VP	E3-VP
account NN	I-NP	I-NP	I-NP	I-NP	I-VP	B3-VP	E2-VP
deficit NN	I-NP	I-NP	E-NP	E-NP	E-VP	E-VP	E-VP
will MD	B-VP	B-VP	I-VP	I-VP	B-VP	B-VP	B-VP
narrow VB	I-VP	I-VP	E-VP	E-VP	E-VP	E-VP	E-VP
to TO	B-PP	B-PP	E-PP	E-PP	S-PP	S-PP	S-PP
only RB	B-NP	B-NP	I-NP	I-NP	B-NP	B-NP	B-NP
# #	I-NP	I-NP	I-NP	I-NP	I-NP	B2-NP	E2-NP
1.8 CD	I-NP	I-NP	E-NP	E-NP	E-NP	E-NP	E-NP
billion CD	B-PP	B-PP	E-PP	E-PP	S-PP	S-PP	S-PP
in IN	B-NP	B-NP	I-NP	I-NP	B-NP	B-NP	B-NP
September NNP	I-NP	I-NP	I-NP	E-NP	E-NP	E-NP	E-NP
. .	O	O	O	O	O	O	O

3.2.Feature template

Graphical models (MEMM and CRFs) are highly dependent on feature templates. For the sake of comparing the effectiveness of different types of features, four different templates are selected for experiments. Context predictions of the current token are sources for feature selection. We firstly introduce atomic features in Table 2 (Ratnaparkhi 1996; Koeling 2000), and four templates are shown in Table 3, Table 4, Table 5 and Table 6. Table 3 and Table 4 shows the templates based on pure lexical and POS information, while Table 5 and Table 6 shows the templates based on mix lexical and POS information. We called them tmpt-1, tmpt-2, tmpt-3 and tmpt-4 in turn.

Table 2: Atomic features

Feature tag	**Feature remark**	**Feature tag**	**Feature remark**
W0	Current word	P0	POS tag of the current word
W-1	The previous word	P-1	POS tag of W-1
W-2	The previous word of W-1	P-2	POS tag of W-2
W1	The next word	P1	POS tag of W1
W2	The next word of W1	P2	POS tag of W2

4. Experiments

We will firstly describe the text chunking data set in detail, then present the chunking performance and discuss it.

Table 3: Features based on pure lexical and POS information

Feature type	Features
Atomic features	W0,W-1,W-2,W1,W2,P0,P-1,P-2,P1,P2
pure features	W-2W-1,W-1W0,W0W1,W1W2,P-2P-1,P-1P0,P0P1,P1P2, P-2P-1P0,P-1P0P1,P0P1P2

Table 4: Features based on pure lexical and POS information

Feature type	Features
Atomic features	W0,W-1,W-2,W1,W2,P0,P-1,P-2,P1,P2
Pure features	W-2W-1,W-1W0,W0W1,W1W2,***W-2W-1W0, W-1W0W1,W0W1W2*** P-2P-1,P-1P0,P0P1,P1P2,P-2P-1P0,P-1P0P1,P0P1P2

Table 5: Features based on mix lexical and POS information

Feature type	Features
Atomic features	W0,W-1,W-2,W1,W2,P0,P-1,P-2,P1,P2
combined features	W-2W-1,W-1W0,W0W1,W1W2, P-2P-1,P-1P0,P0P1,P1P2,P-2P-1P0,P-1P0P1,P0P1P2, P-1W-1,P0W0,P-1P0W-1,P-1P0W0,P-1W-1W0,P0W-1W0,P-1P0P1W0

Table 6: Features based on mix lexical and POS information

Feature type	Features
Atomic features	W0,W-1,W-2,W1,W2,P0,P-1,P-2,P1,P2
combined features	W-2W-1,W-1W0,W0W1,W1W2,***W-2W-1W0, W-1W0W1,W0W1W2*** P-2P-1,P-1P0,P0P1,P1P2,P-2P-1P0,P-1P0P1,P0P1P2, P-1W-1,P0W0,P-1P0W-1,P-1P0W0,P-1W-1W0,P0W-1W0, P-1P0P1W0

4.1.Experimental Setting

Our data set comes from CoNLL-2000 shared task (Tjong Kim Sang and Buchholz, 20001). In this data set, the total of 10 base phrase classes (NP, VP, PP, ADJP, ADVP, CONJP, INITJ, LST, PTR, SBAR) are annotated. This data set consists of 4 sections (15-18) of the WSJ part of the Penn Tree bank for the training data and one section (20) for the test data. I In order to show the relationship between M3Ns and the data set size, we split the CoNLL-2000 training data set into parts with different size: 20%, 40%, 60%, 80% and 100%. For the kernel function, we use the linear kernel function with margin parameter C=1.

In the text chunking task, three rates are usually used to measure the performance of the systems. They are precision P, recall R and F_β.

$$P = \frac{\#\ of\ correct\ proposed\ chunk}{\#\ of\ proposed\ chunk} \quad R = \frac{\#\ of\ correct\ proposed\ chunk}{\#\ of\ corect\ chunk} \quad F_\beta = \frac{(\beta^2+1)RP}{\beta^2 R + P}(\beta = 1)$$

4.2.Experimental Results

In the experiment, we compare the performance of different representations and different templates. We also investigate the affects of different sizes of training data to validate the generalization ability of M3Ns.

Firstly, we use tmpt-3 with different **Inside/Outside** templates. Table 7 shows the results of M3Ns on the whole data set. We can see that there is no great difference between them. Secondly, we compared the performance of different templates with representation IOB2. Table 8 shows the experiment result that templates based on mixed lexical and POS information (tmp-3 and tmpt-4) are more suitable than templates based purely on lexical or POS information (tmpt-1 and tmpt-4). Besides, the second-order lexical features such as W-2W-1W0 are not always good. At last, in order to validate the high generalization ability of M3Ns, we compared

[1] http://www.cnts.ua.ac.be/conll2000/chunking/

the performance of M3Ns and CRFs[2] on the same training data sets of different sizes. Figure 1 shows the experiment result that the M3Ns achieve better performance and the curve of M3Ns goes more smoothly.

Table 7: Results of different chunk representations on whole data set

	Inside/Outside(tmpt-3)			
	IOB1	IOB2	IOE1	IOE2
precise	93.57	**93.72**	93.42	93.56
recall	93.38	**93.54**	93.40	93.54
F1	93.48	**93.63**	93.45	93.55

Table 8: Results of different templates on whole data set

	IOB2			
	tmpt-1	tmpt-2	tmpt-3	tmpt-4
precise	93.60	93.67	93.72	**93.74**
recall	93.21	93.30	**93.54**	93.41
F1	93.40	93.48	**93.63**	93.58

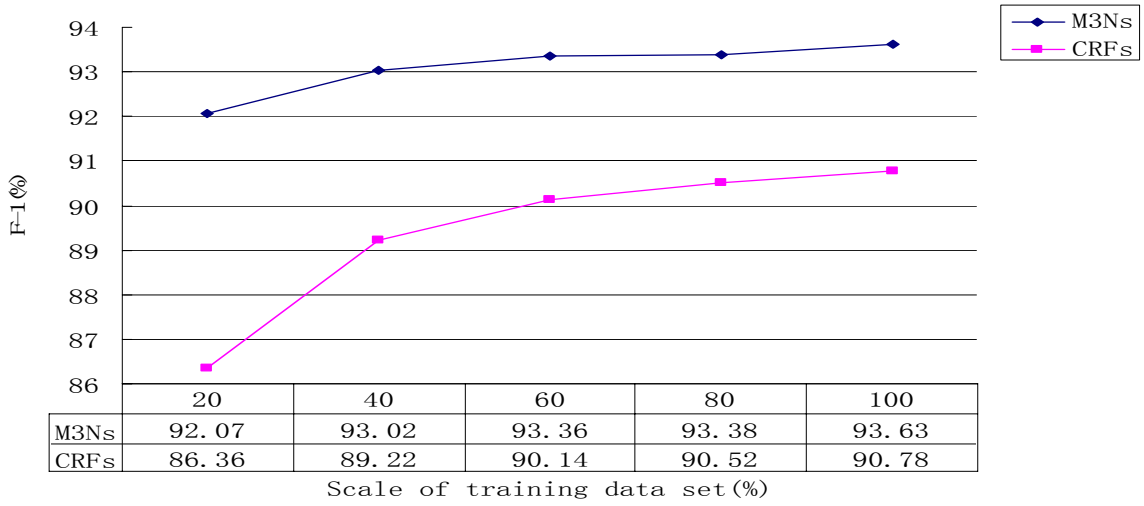

Figure 1: results for CoNLL-2000 training data sets of different size using IOB2 representation

4.3.Comparison with Related Works

In this section, we compare our results with eleven systems in CONLL-2000 (Tjong Kim Sang et al., 2000). Table 9 shows the performance of our system and systems in CONLL-2000.

Table 9: Comparison with systems in CONLL-2000

systems	precision	recall	F	systems	precision	recall	F
Our	**93.74%**	**93.54%**	**93.63**	Koe00	92.08%	91.86%	91.97%
KM00	93.13%	93.51%	93.48%	Osb00	91.65%	92.23%	91.94%
Hal00	93.13%	93.51%	93.32%	VB00	91.05%	92.03%	91.54%
TKS00	94.04%	91.00%	92.50%	PMP00	90.63%	89.65%	90.14%
ZST00	91.99%	92.25%	92.12%	John00	96.24%	88.25%	87.23%
Dej00	91.87%	92.31%	92.09%	VD00	88.82%	82.91%	85.76%

Clearly, M3Ns performed better than all systems in CONLL-2000. Especially, better than single-algorithm systems: Rule-based systems by Villain and Day, Johansson, and Dejeanbetter; Memory-based systems by Veenstra and Vanden Bosch; and Statistical systems by Pla, Molina and Prieto, Osborne, Koeling, Zhou, and Tey and Su.

Here, we should mention that some successful systems combined (Taku Kudoh and Yuji Matsumoto, 2001) or features (Zhang et al., 02) enhanced have been better than ours. However,

[2] http://crfpp.sourceforge.net/

it is not a fair comparison to our system since it is reasonable to believe that we can achieve appreciable improvement in the similar approaches.

5. Summary

In this paper, we introduce a text chunking system based on Max-Margin Markov Networks. Since M3Ns make full use of correlations in data like CRFs, they can achieve good performance using the same features of CRFs. Furthermore, due to the theoretical generalization guarantee, M3Ns also have special error toleration ability. In our experiments, we have shown that M3Ns perform better than CRFs with high generalization ability. The success of M3Ns in text chunking suggests that the approach might be applicable to other NLP problems such as Part Of Speech (POS) and Named Entity Recognition (NER).

Reference

GuoDong Zhou, Jian Su and TongGuan Tey, Hybrid Text Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.

Erik F. Tjong Kim Sang and Sabine Buchholz, Introduction to the CoNLL-2000 Shared Task: Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.

Rob Koeling, Chunking with maximum entropy models, Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning, September 13-14, 2000, Lisbon, Portugal.

John D. Lafferty, Andrew McCallum and Fernando C. N. Pereira. 2001. Conditional Random Field: Probabilistic models for segmenting and labeling sequence data [A]. In: ICML-18 [C], pp.282-289. June 28-July 01, 2001.

Nianwen Xue and Libin Shen. Chinese word segmentation as LMR tagging [A]. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing [C], pp.176-179, Sapporo, Japan: July 11-12, 2003.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation -based learning. In Proceedings of the 3rd Workshop on Very Large Corpora,pages 88–94.

Kiyotaka Uchimoto, Qing Ma, Masaki Murata, Hiromi Ozaku, and Hitoshi Isahara. 2000. Named Entity Extraction Basedon AMaximum Entropy Model and Transformation Rules. In Processing of the ACL2000.

Collins, M. (2002a). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In Proceedings of EMNLP 2002.

Michael I. Jordan, editor. Learning in Graphical Models. MIT press, Cambridge, MA, 1998.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. Machine Learning, 20(3):273–297, 1995.

V. Vapnik. The Nature of Statistical Learning Theory. Statistics for Engineering and Information Science. Springer, New York, NY, 1999.

Yoav Freund and Robert E.Schapire. 1996. Experiments with a new boosting algorithm. In International Conference on Machine Learning(ICML), pages 148–146.

Ratnaparkhi, A.,“A Maximum Entropy Model for Part-Of-Speech Tagging,” In Proceedings of EMNLP’1996, New Brunswick, New Jersey, USA, 1996, pp. 133-142.

Tong Zhang, Fred Damerau and David Johnson, Text Chunking based on a Generalization of Winnow. In Journal of Machine Learning Research, volume 2 (March), 2002, pp. 615-637.

Taku Kudoh and Yuji Matsumoto, Chunking with Support Vector Machines, In: "Proceedings of NAACL 2001", Pittsburgh, PA, USA, 2001.